Tapping the implicit information for the PS to DS conversion of the Chinese Treebank

Nianwen Xue
University of Colorado
Center for Computational Linguistics and Education Research

Abstract

We examine the linguistic adequacy of dependency structure annotation automatically converted from phrase structure treebanks with the head table approach and show this method is far from satisfactory. We propose an alternative approach that better exploits the implicit information in the phrase structure and show these two approaches only agree 60.6% of the time when evaluated against the Chinese Treebank.

1 Introduction

The dominance of data-driven approaches to natural language parsing has spurred the development of a large number of treebanks in a variety of different languages (Hajič et al., 2003; Brants, Skut, and Uszkoreit, 2003; Abeillé, Clément, and Toussenel, 2003; Kurohashi and Nagao, 2003; Han et al., 2002; Marciniak et al., 2003; Moreno et al., 2003; Oflazer et al., 2003; Xue et al., 2005). These treebanks are in turn annotated with a wide range of representation schemes reflecting the diversity of languages and linguistic traditions. Generally, treebanks for languages that have a freer word order tend to adopt a *dependency structure* representation (e.g., the Prague Dependency Treebank (Hajič et al., 2000)), emphasizing the grammatical relation between the head and its dependents, while treebanks for languages with a more rigid word order tend to use a *phrase structure* representation, stressing the hierarchical organization of the constituents of a sentence. In a dependency structure treebank, typical dependency categories are subject, object, etc., which imply the role the dependent plays with regard to its head. In a phrase structure representation, the phrasal category of a constituent generally embodies the distributional properties of the constituent in a larger structure. The two representation schemes do not necessarily preclude one from the other in an annotation framework. While the Penn English Treebank (Marcus, Santorini, and Marcinkiewicz, 1993) is a typical phrase structure treebank, it also has elements that represent dependency relations. In addition to phrasal labels, it also has functional

De Smedt, K., Hajič, J. and Kübler, S. (Eds.)
Proceedings of the Sixth International Workshop
on Treebanks and Linguistic Theories (2007)

tags like SBJ (subject), OBJ (object), TMP (temporal) and LOC (location) that represent the relation between constituents. In addition, it uses empty categories and co-indexation mechanisms to represent long-distance dependencies. However, the representation of dependencies in Penn Treebank is incomplete. Not all dependencies are explicitly represented if represented at all. More recent treebanks like the Tiger Treebank (Brants et al., 2002) and TüBa-D/Z treebank (Telljohann et al., 2005) for German seek to explicitly represent both constituent and dependency structures by labeling both nodes and edges in the syntactic tree.

In the absence of constituent or dependency structures for a particular treebank, NLP researchers, especially natural language parser developers, have resorted to automatic conversions from one representation scheme to another to get the necessary data to train and test their parsing algorithms. Automatic conversion has been done for both directions and different issues arise. One main issue in the conversion from the dependency structure to the constituent structure is the indeterminency in the choice of a phrasal category given a dependency relation, the level and position of attachment of a dependent in the constituency structure, as dependency relations typically do not encode such information (Xia and Palmer, 2001). All dependents of the same head are considered equal in their closeness to the head while a phrase structure representation scheme often makes the distinction between dependents that are more "important" to the head and the ones that are less important. An example is the argument/adjunction distinction, where the argument is attached at a level that is "closer" to the head than the adjuncts. In languages where non-projective dependencies are abundant, there is also the need to transform them into a projective structure that is easy to manipulate and process.

The conversion from phrase structure to dependency structure, even though used more often, has received less scrutiny because it is considered more straightforward. The automatically converted data from phrase structure treebanks such as the Penn English Treebank (Marcus, Santorini, and Marcinkiewicz, 1993) and the Chinese Treebank (Xue et al., 2005) has been used to train and test state-of-the-art parsing algorithms (Nivre et al., 2007a; McDonald, 2006), presumably because there is less concern in the reliability of such automatically converted data. To a certain extent, the conversion from phrase structure to dependency structure is in fact easier because most of the indeterminencies found in the dependency to phrase structure conversion do not exist. As noted above, most phrase structure treebanks provides some dependency annotation that can be used to mark dependency relations, even though such information is implicit and incomplete. Since a dependency representation scheme does not use the level of attachment as a way to represent linguistic information, there is no indeterminency in the phrase structure to dependency structure conversion in this regard. All dependents are directly linked to the head, in a flat structure.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

Finding the head, therefore, is of paramount importance when a phrase structure is converted to a dependency structure. While in a phrase structure the head does not have to be explicitly annotated, in a dependency structure knowing the head is essential.

Most of the phrase structure to dependency structure conversion algorithms use a head table to find the head of a phrase. Such head tables originated from statistical parsing literature (Collins, 1999) and consist of a list of rules defined in relation to a phrasal category. The head table approach provides a simple, heuristic way to find the head, and is widely used in phrase structure to dependency structure conversion. However, there has been thus far very little evaluation as to whether the heads found in this manner are linguistically justified. Given the increasing popularity of data-driven dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007b), it is timely to examine the reliability of the automatically converted data to ensure the healthy evolution of the dependency parsing research. This paper reports the result of a comparison of two phrase structure to dependency structure conversion strategies. One strategy uses the head table approach, and another uses a more structural approach that first determines the grammatical relations between the constituents. For the former we use a publicly available phrase structure to dependency structure conversion utility called Penn2Malt [1]. For the latter approach we use a tool we developed in house. An evaluation on the Chinese Treebank showed that the two approaches agree only 60.6% of the time in terms of unlabeled dependencies. Without prejudging which of these two approaches provides linguistically more plausible structures, this result shows that making the correct linguistic determination in the dependency structure representation is crucial for the health of dependency structure parsing. Without sound linguistic underpinnings in the data annotated with dependency structures, the parsers cannot be properly evaluated.

This paper is organized as follows. In section 2, we demonstrate how the head table approach works and show some of the problematic dependency structures produced by this approach. In Section 3 we provide an alternative approach that takes advantage of some of the structural information represented in the Chinese Treebank to get more accurate phrase structure to dependency structure conversion. Section 4 concludes.

2 The head table approach

A head table is a list of rules that can be used to guide the search for the head of a constituent in a phrase structure tree. A head rule is defined relative to a phrasal category. That is, different rules apply for constituents of different phrasal categories. For example, there is one rule for NPs and

[1] The tool can be found at http://w3.msi.vxu.se/ nivre/research/Penn2Malt.thml

another for VPs. Typically, a head rule specifies the order as well as the direction in which the head can be located, given the phrasal category. For example, the rule used by Penn2Malt (as well as many other PS to DS conversion algorithms) to find the head of the VP is:

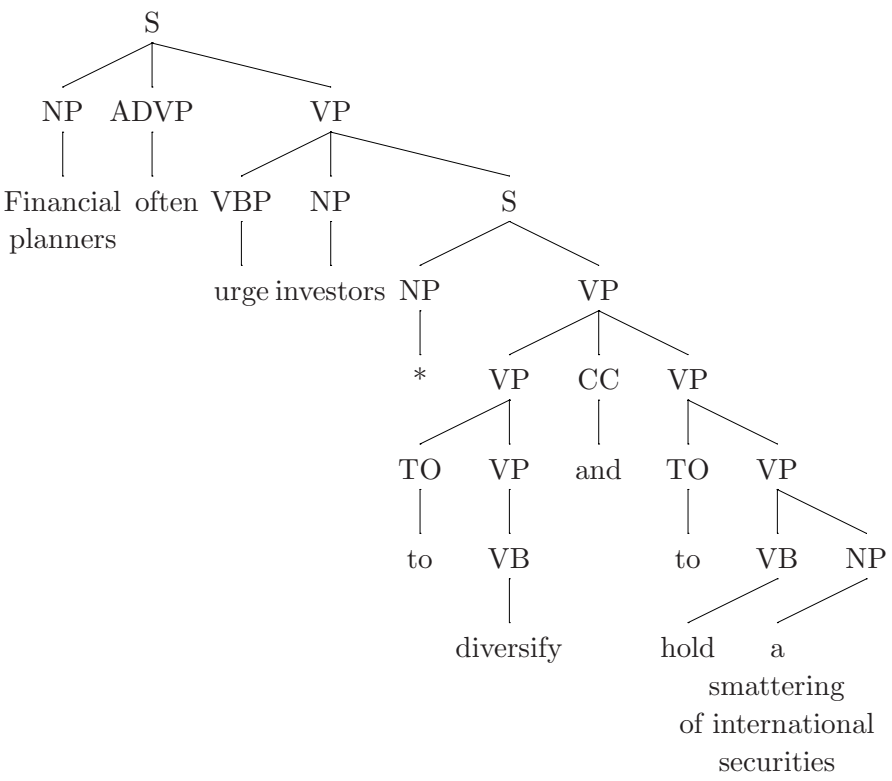
VP *l VBD;l VBN;l MD;l VBZ;l VB;l VBG;l VBP;l VP;l ADJP;l NN;l NNS;l NP;l*

The rule basically states that to find the head of a VP constituent, search for a constituent labeled as one of VBD, VBN, MD, VBZ, VB, VBG, VBP VP ADJP, NN, NNS, NP, in that order, from left to right. The first match is the head of the VP constituent. In a dependency structure, the head always has to be a lexical head. So in cases where the head found is a non-terminal node in the phrase structure tree, the PS to DS conversion algorithm would recursively try to find the head of this non-terminal node until it finds the lexical head. For this particular VP rule, if the head found is a VP, ADJP, or NP, then the conversion algorithm would recursively try to locate their head until finding a lexical head. The head rule approach works as long as there is exactly one head for each constituent. (1) is an example from the Penn Treebank and (2) is its dependency structure representation that Penn2Malt produces given the phrase structure representation in (1) as input. As can be seen, the conversion algorithm correctly identifies the lexical head of the VP, "urge", as the head of "planners" (subject), "often" (adverbial modifier), "investors" (object). It also correctly identifies "planners" as the head of "Financial" (adjectival modifier), "securities" as the head of "international" (adjectival modifier), "of" as the head of "securities" (noun complement to preposition). It is also reasonable to assume that "smattering" is the syntactic head the determiner "a" and the preposition "of".

The head table approach runs into trouble when there is a coordination structure. For example, Penn2Malt identifies the lexical head of the first conjunct, "diversify", as the head of the conjoined VP and also the head of the embedded clause S. The coordination conjunction "and" and the lexical head of the other conjunct "hold" are identified as dependents of "diversify". Linguistically there are at least two problems with this. First of all, this precludes the linguistic dependency between "urge" and "hold". The financial advisers urged investors to do two things: "diversify" and "hold a smattering of international securities", not just one. Second, treating the coordination conjunction "and" and the other conjunct "hold" similarly, both as modifiers of the "diversify" is linguistically odd as well.

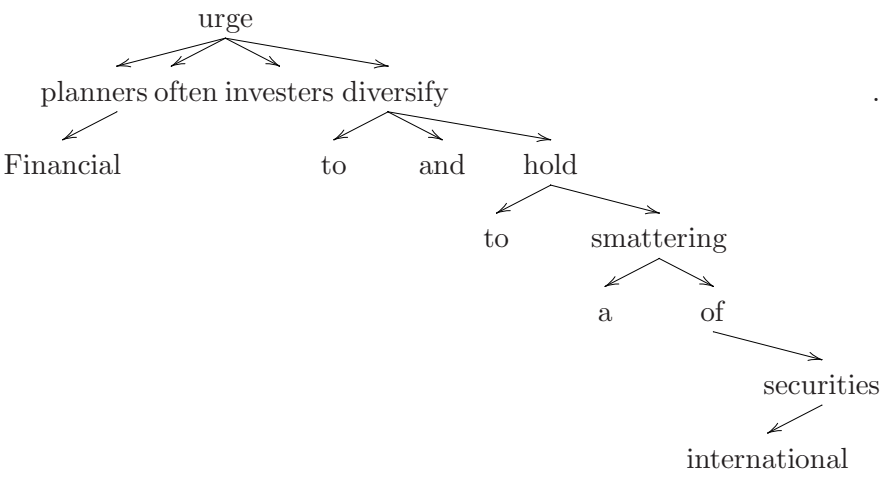
ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

(1)



"Financial planners often urge investors to diversify and to hold a smattering of international securities."

(2)

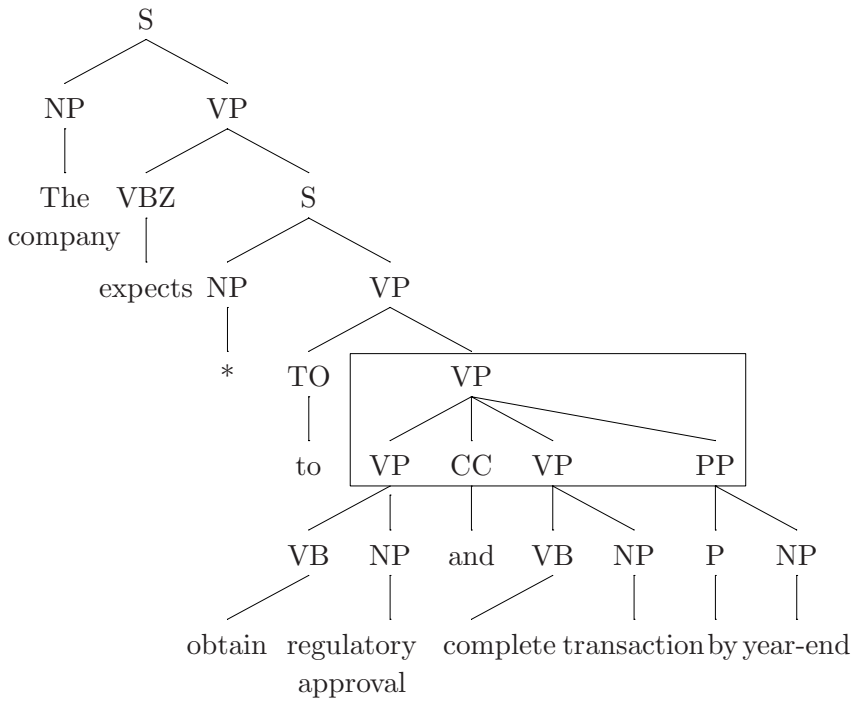


.

A solution to this problem would require recognizing that this VP is a coordination structure and treating coordinating structures differently than other VP structures where the grammatical relation between the head and its dependent is one of modification or complementation. In (1), there is

enough information in the phrase structure to allow an algorithmic determination that this is a coordination structure, given the presence of a coordination conjunction (CC) "and" and two VP conjuncts that share the phrasal category of the parent constituent. However, this is not consistently the case in the Penn Treebank. There are many cases in the Penn Treebank where a coordination conjunction is attached at the same level as a modifier. For example, in (3), the coordination conjunction "and" conjoins the VP "obtain regulatory approval" i with the VP "complete transaction", but the PP "by year-end" is attached at the same level as a modifier of both. This makes the automatic detection of coordination structures difficult, if not outright impossible.

(3)



"The company expects to obtain regulatory approval and complete transaction by year-end."

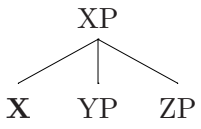
3 A Structural approach to PS to DS conversion

The influence of the Penn Treebank on the development of the Chinese Treebank is obvious. The Chinese Treebank essentially adopted the Penn Treebank annotation scheme and used the same grammatical devices to represent syntactic relations. Like the Penn Treebank, the Chinese Treebank uses a combination of configurational and non-configurational mechanisms. It uses phrasal categories that represent the distributional properties

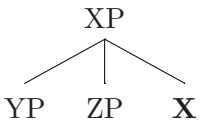
ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

of phrases as well as functional tags that represent grammatical relations between constituents. It also uses empty categories and indices to represent long-distance dependencies. However, there are also important differences between the two. One such difference is that the Chinese Treebank enforces the "one grammatical relation per bracket" policy so that each constituent in a syntactic tree falls into one of the three primitive grammatical relations in (4). The grammatical relation for each constituent is one of complementation where a terminal node is the head (which can be head-initial (4a) or head-final (4a')) taking a non-terminal node as its complement, adjunction where a non-terminal head has non-terminal modifiers, or coordination where constituents are conjoined with one or more coordination conjunctions. This means that the Chinese Treebank would not allow a structure like (3) where different types of grammatical relations co-exist within one constituent. It would force the PP "by year-end" to be attached at a different (higher) level than the conjoined VPs "obtain regulatory approval" and "complete transaction".

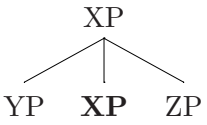
(4)



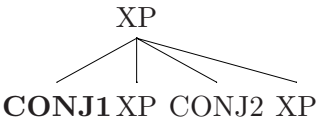
a. Head-initial complementation



a'. Head-final complementation



b. Adjunction



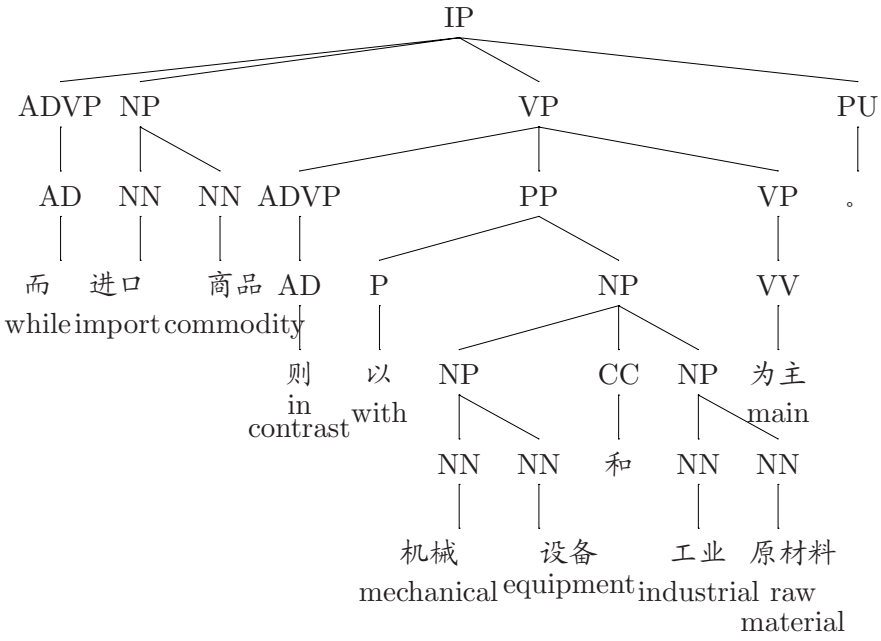
c. Coordination

This approach draws mixed reviews from NLP researchers who are used to working with the Penn English Treebank annotation style (Levy and Manning, 2003). On the one hand, it forces more structures for the same sentence and presumably makes parsing more difficult. On the other hand, it makes the grammatical relations within each constituent more uniform. It opens the door for alternative approaches to PS-DS conversion. For example, this makes it possible to algorithmically differentiate coordination structures from other constituent types, among other things. Once the grammatical relation is identified, then one can follow different procedures to identify the head for different types of constituents. For complementation structures, one can just try to find the first (non-punctuation mark) terminal, left to right or right to left, based on the phrasal category of the constituent. For adjunction

structures, the head is almost always the final non-terminal child that has the same phrasal category as the parent For these two types of constituent structures, the procedure for identifying the head is not that different from the head table approach. The empirical results from a comparison of our approach and the head table also support this observation.

The difference is with coordination structures. Once a coordination structure is identified, we make the following determination as to what the head is. (i) we assume that the head is the coordination conjunction, not the conjunct, following the Prague Dependency Treebank, (ii) we stipulate that the first coordination conjunction is the head if there are multiple coordination conjunctions, and (iii) in the rare cases where there i no conjunction, the first conjunct is designated as the head. Running the conversion algorithm outlined here on the first 250K of the Chinese Treebank 6.0 (files chtb_0001.fid to chtb_0931.fid), and comparing its output against the output of Penn2Malt, we found that the total agreement in terms of unlabeled dependencies is an alarming lowly 60.6%! (5) is a sentence from the Chinese Treebank, (6) is the output of the our conversion procedure, and (7) is the output of Penn2Malt. Notice that difference is not only with the coordination structure, but also dependencies links interacting with the coordination structure.

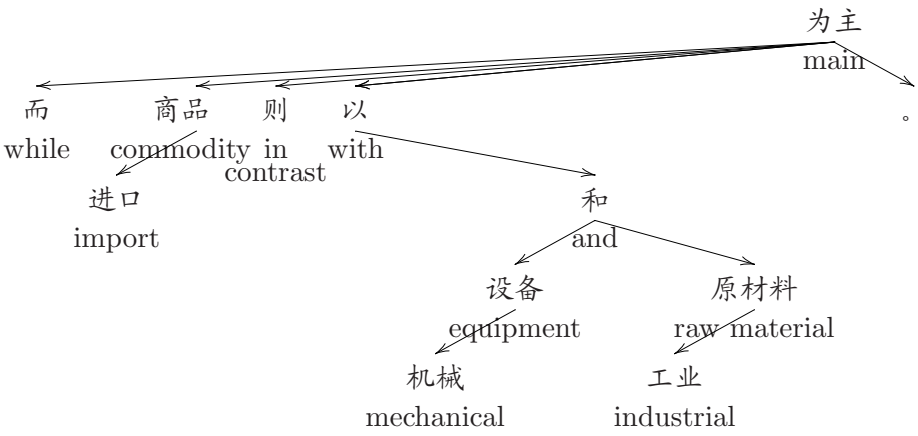
(5)



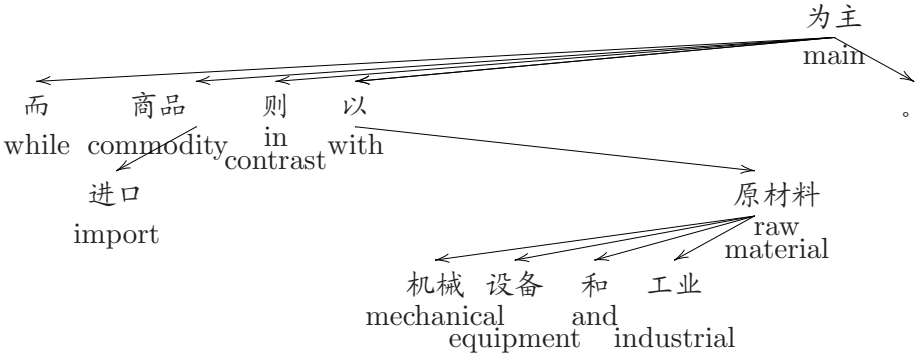
"In contrast the imported commodities are mainly mechanical equipment and industrial raw material."

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

(6)



(7)



Designating the coordination conjunction as the head of a coordination structure is not free from problems. The difficulty in representing coordination structures in general originates from the dual status of the head (Corbett, Fraser, and McGlashan, 1993; Meyers, 1995), as the thematic head and the functor. Modification and complementation structures do not pose similar problems because these two roles converge. That is, their head function as both the thematic head and the functor. In a coordination structure, the thematic head, which determines its external selectional restrictions, should be a conjunct or the collection of conjuncts. The functor, which is the glue that holds the coordination structure together, is the conjunction. Taking the functor as the head makes the coordination structure-internal dependencies linguistically plausible, at the expense of the dependency between coordination structure as a whole and its head. This problem is generally addressed by treating this external dependency as transparent. That is, any dependency that holds between the functor and its head is transferred to the immediate dependents of the functor.

4 Conclusion

We examined the linguistic adequacy of the dependency structure annotation automatically converted from phrase structure treebanks and found that the coordination structures in particular are not properly converted with the widely used head table approach. We then proposed an alternative approach that better exploits the structural information in the Chinese Treebank. An evaluation on a 250K-word portion of the Chinese Treebank shows that these two approaches agree only 60.6% in terms of unlabeled dependency. Without prejudging which conversion algorithm is superior, this result shows that it is important to make a sound linguistic determination with regard to the coordination structures in order for the automatically converted data to be used to train and test dependency parsing algorithms.

References

Abeillé, Anne, Lionel Clément, and François Toussenel. 2003. Building a Treebank for French. In Anne Abeillé, editor, *Treebanks: Building and Using Annotated Corpora*. Kluwer Academic Publishers.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Brants, Thorsten, Wojciech Skut, and Hans Uszkoreit. 2003. Syntactic Annotation of a German Newspaper Corpus. In Anne Abeillé, editor, *Treebanks: Building and Using Annotated Corpora*. Kluwer Academic Publishers.

Buchholz, Sabine and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City, June. Association for Computational Linguistics.

Collins, Michael. 1999. *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Corbett, G., N. M. Fraser, and S. McGlashan. 1993. *Heads in Grammatical Theory*. Cambridge University Press.

Hajič, Jan, Alena Böhmová, Eva Hajičová, and Barbora Vidová-Hladká. 2000. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In A. Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*. Amsterdam:Kluwer, pages 103–127.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

Hajič, Jan, Alena Böhmová, Eva Hajicová, and Barbora Hladká. 2003. The Prague Dependency Treebank: A Three Level Annotation Scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Annotated Corpora*. Kluwer Academic Publishers.

Han, Chunghye, Narae Han, Eonsuk Ko, and Martha Palmer. 2002. Korean treebank: Development and evaluation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, Las Palmas, Spain.

Kurohashi, Sadao and Makato Nagao. 2003. Building a Japanese Parsed Corpus While Improving the Parsing System. In Anne Abeillé, editor, *Treebanks: Building and Using Annotated Corpora*. Kluwer Academic Publishers.

Levy, Roger and Christopher Manning. 2003. Is it Harder to Parse Chinese, or the Chinese Treebank. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan.

Marciniak, M., A. Mykowiecka, A. Przepiorkowski, and A. Kupsc. 2003. Construction of an HPSG treebank for Polish. In Anne Abeillé, editor, *Treebanks: Building and Using Annotated Corpora*. Kluwer Academic Publishers.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

McDonald, Ryan. 2006. *Discriminative Training and Spanning Tree Algorithms for Dependency Parsing*. Ph.D. thesis, University of Pennsylvania.

Meyers, Adam. 1995. The NP analysis of NP. In *Proceedings of the 31st Regional Meeting of the Chicago Linguistic Society*, pages 329–342.

Moreno, Antonio, Susana Lopez, Fernando Sanchez, and Ralph Grishman. 2003. Developing a Syntactic Annotation Scheme and Tools for a Spanish treebank. In Anne Abeillé, editor, *Treebanks: Building and Using Annotated Corpora*. Kluwer Academic Publishers.

Nivre, J., J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. 2007a. MaltParser: A Language-independent system for data-driven dependeny parsing. *Natural Language Engineering*, 13(2):95–135.

Nivre, Joakim, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007b. The CoNLL 2007 Shared

Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Oflazer, Kemal, Bilge Say, Dilek Zeynep Hakkani-T*ü*r, and Gökhan T*ü*r. 2003. Developing a Morphologically and Syntactically Annotated Treebank Corpus for Turkish. In Anne Abeillé, editor, *Treebanks: Building and Using Annotated Corpora*. Kluwer Academic Publishers.

Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, and Heike Zinsmeiste. 2005. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z).

Xia, Fei and Martha Palmer. 2001. Converting Dependency Structures to Phrase Structures. In James Allan, editor, *First International Conference on Human Language Technology Research*. Morgan Kaufmann, pages 61–65.

Xue, Nianwen, Fei Xia, Fu dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207–238.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476